Electronic California Documents Pilot Project (Final Report)

I. Overview

The Cataloging and Metadata Common Interest Group (CAMCIG) of the University of California began discussions in Spring 2008 about new approaches to California electronic document cataloging in response to a looming cut to UC's Shared Cataloging Program (SCP) budget. SCP had been responsible for the cataloging of electronic publications for a list of California state agencies (a selected list developed by the UC Government Information Librarians group), and distributing those records to the ten UC campuses for inclusion in their local OPACs as well as the union catalog, Melvyl. The $48,000 budget cut, effective July 1, 2008, reduced SCP staff levels and made continuing this work impossible. Recognizing that the critical work of electronic state document cataloging and record distribution must continue, CAMCIG approved two complementary proposals that would share the intellectual work through collaborative effort and take advantage of automated processes.

Printed state documents that come to UC libraries are not a part of this effort, and are still handled separately by staff in those depository libraries.

II. **Automated Record Harvesting and Distribution** (also called the SCP Proposal)

Based on a new search capability within OCLC's WorldCat, SCP staff proposed that a regular, monthly search could be implemented--with little or no manual intervention--that would gather electronic California documents records that had been created or updated within a specific date range. This automated search would include any records created through the Shared Record Creation Proposal (see section III below). Such an automated search would not be perfect: some records would be missed, and individual records would not have their content evaluated (e.g., broken links would not be fixed and PURLs would not be created for any links, contrary to standard SCP practice). Because the process is fully automated, SCP staff could do this quickly, and the records could be distributed to local OPACs (and from them, to Melvyl) as part of SCP's regular weekly distribution process.

The search criteria and details of the processing workflow can be found on the SCP Web site at http://www.cdlib.org/inside/projects/scp/caldocsonline.pdf.

Two UCLA staff members, early on, volunteered to do some analysis of the monthly harvested packages so that we can understand better the quality tradeoffs we are making by not manually examining the records.

In the end, though, the harvesting routine proved to be much more complicated and time-consuming than originally anticipated. For SCP catalogers the process of harvesting was not a positive experience. The harvesting was very labor-intensive, in particular, the time spent fine-tuning the process, troubleshooting and communicating. The original harvest algorithm was too broad, bringing in large amounts of records with broken links and other issues. More recent changes to the algorithm may help, but we need to monitor this further before deciding whether the changes are working or not. As UCSD's report points out, the dramatic change to a fully-automated process (with no possibility for manual review) has been a struggle for all. Yes, UC should be proud of the work involved in inventing and reinventing this new model.

Some concerns/questions expressed:

- Have not had enough time to see how the newly revisited harvest is working
- In the early stages of the pilot, record loading was performed manually at campuses; catalogers were overwhelmed initially when the harvest sent hundreds of records in a single month; process was very labor intensive; new changes to the harvesting algorithm have greatly reduced the number of records.
- Prior to the recent changes in the harvesting algorithm, catalogers saw little value in the harvesting routine; recent comments from the campuses indicate the new algorithm is working much better.
- Early harvest had high number of errors (50% of records deleted on one campus); recent harvest model had 22% of records with broken links
- Few new titles coming in harvested file
- CalDocs have a higher failure rate (14.4%) than the other URLs in the library catalog

III. **Shared Record Creation** (also called the Berkeley Proposal)

UC Berkeley staff developed a proposal to divide the GILS agency list among five campuses, with each taking responsibility for monitoring those agencies for new electronic publications, selecting publications for cataloging, checking WorldCat to see if copy exists, and creating a new record if no copy is found. Five campuses (Berkeley, Davis, Irvine, Los Angeles and San Diego) made the commitment to participate, and the agencies each took responsibility for are shown in Appendix A. CAMCIG agreed that this would be a pilot project subject to evaluation in March 2009. Campuses began original cataloging in September 2008, and staff are tracking data in terms of both selection and cataloging that will be used in project assessment.

Each campus handled their own workflow, and because we assumed that each campus was fully capable of handling original cataloging, there were no written procedures to coordinate this effort. We expected the five campuses to handle serials, monographs, and integrating resources. It was left to each campus to decide whether or not to upgrade substandard copy. CAMCIG agreed not to mandate LC or CSL classification for these electronic materials. We did not expect to catalog all publications of an agency. The need for a sustainable digital preservation program for electronic California documents was expressed for a variety of good reasons. There were also some questions about single vs. separate record techniques, including the official cataloging policy of the State Library, and how print workflows mesh (or do not mesh) with electronic.

Effective February 1, 2009, CAMCIG, following the recent policy change by the GPO, approved the use of separate records for government document monographs. Historically, UC followed a separate record policy for all monographs except government documents. A change in UC policy will mean that all monographs will be cataloged using the separate record approach. Faced with staffing shortages and budget reductions, separate records for government document monographs should facilitate automatic loading of records. Although the State Library has historically used the single record approach for monographs, CSL catalogers are now working with UC catalogers to come up with a macro that would allow for a relatively quick and easy production of a separate record for online resources. Beginning in February 2009 the State Library will be creating these separate records for newly cataloged state publications.

Campuses were asked to provide the following information:

- Total number of titles selected
- Total number of hours spent on selection
- Total number of titles cataloged
- Total number of titles already cataloged in WorldCat
- Total number of hours spent on cataloging
- Comments on how well the original cataloging process worked

Specific campus responses can be found in Appendix B.

The general consensus from the campuses indicated most campuses were happy with this component of the pilot. UCSD suspects that we are now generating more cataloging than SCP did under the previous model which is a positive trend for access and for users. The recent decision from CAMCIG to provide separate records for electronic government document monographs should facilitate the quick creation of monographic records. UCSD has developed a macro that facilitates this process and could be shared with the other campuses.

Some concerns/questions expressed:

- Are we all focusing on NEW content and NOT searching for retrospective titles?

- It seems we're not finding many resources that seem worthwhile of selection & cataloging from the agencies assigned to UCLA. Are there other campuses that are feeling overburdened with original CalDocs cataloging, and should we consider redistributing the agencies to make for a more evenly distributed workload? (UCLA)

IV. Selection Process

As mentioned, each campus determined locally how their selection process would work in conjunction with their catalogers.

- Hard to locate documents on some agency websites
- Checking whether a document has been cataloged or not is time-consuming
- Little interaction between selector and catalogers
- Selector finds that the harvest is not capturing all of the significant titles found on the recent California State Publications list
- Takes 6-8 hours to review each agency the first time
- Necessary to check each title in WorldCat during selection routine since the cataloger will need to recheck in WorldCat?
- Our selector provided criteria for selection rather than a title list; this worked well
- It seems we're not finding many resources that seem worthwhile of selection

V. Archiving Issues/Activities

A. ContentDM trial at UCB

For the last three months of the pilot UC Berkeley catalogers tested ContentDM using an existing license involving several affiliated libraries (i.e. Water Resources Center, Institute of Governmental Studies, Law and Transporation Studies) on the Berkeley campus. The affiliated libraries agreed to allow CalDoc catalogers to use their instance of ContentDM so the titles could be archived at the point of cataloging. The main reasoning for testing this service was to gather pertinent information which might be used during future discussions related to archiving documents. Based on comments from catalogers during this testing, the process went well. It took an average of 10 minutes to archive each document. For the next six month period, the Berkeley catalogers recommend testing CDL’s Web Archiving Service (WAS) and then compare. It is possible that archiving at point of cataloging projects could increase depending upon Next Gen Technical Services initiatives, so this initial test may help information gathering for future projects.

B. Discussions with GILS and PAG

Early in the pilot project CAMCIG members voiced concerns about pursuing further discussions and activities with respect to archiving the CalDocs. It was felt this really was an issue for the GILS or the PAG group to pursue (not CAMCIG) CAMCIG members felt our efforts were needed in the cataloging of the agency titles. A. Barone contacted

both the chair from PAG and GILS. Ultimately, it was decided that GILS was the group to pursue this issue. Recently, GILS members agreed we should be archiving California documents we catalog. GILS members are willing to use Web Archiving Service & exploring movement of items to WAS from other systems like Content DM. Those campuses using ContentDM should continue to do so. Ultimately may want to move our objects to WAS. Tracy Seneca, CDL WAS manager, noted from a preservation standpoint, interoperability is important. She'd like us to look to WAS to fulfill our needs, but is not aiming for an either/or choice.

VI. Recommendations:

1. Several campuses, as well as the GILS members, feel all documents cataloged by UC libraries should be digitally archived. As mentioned earlier, the GILS group is pursuing this issue and looking to the Web Archiving Service (WAS).

Next Step:

Berkeley has contacted Tracy Seneca at CDL asking her to set the Berkeley catalogers up for a six month Web Archiving Service (WAS) archiving pilot. Berkeley can start archiving effective June 1st, 2009. Through August 2009, UCB will test WAS for archiving at point of cataloging for all the CA documents that our selector requests (in our agency list). Depending on the outcome of the UCB WAS test, CAMCIG recommends testing for an additional six month project wherein all of the five CA Doc cataloging agencies pilot the archive at point of cataloging method (assuming the Berkeley experience is positive). The archiving we're testing out would be ONLY for those documents we do original cataloging on. We would not archive for copy cataloged titles nor old titles (those with existing records in our catalogs) These two things would be out of scope.
It's possible over time, if this works out we could begin to include archiving for all titles we copy catalog.

Phase 1 UCB tests WAS
Phase 2 - all 5 campuses test archiving at point of cataloging for original records using either WAS or ContentDM
Phase 3 - add archiving at point of cataloging for copy cataloged titles

2. CAMCIG members agreed that it is too soon to fully understand what impact the recent changes to the SCP harvest for copy cataloging CA Doc titles has made.

Next Step:

Continue to monitor the harvesting process and assess its effectiveness; provide answers with regard to the following:

- If harvesting process proves to be too problematic, are we willing to let it go

and possibly pursue alternate means of CalDoc record distribution?

- Are we satisfied with the harvest process at this point? What is "good enough"; when does time spent outweigh the usefulness; user assessment process needed?
- Since those records are in OCLC already, UC library users should be able to access those resources via Next Gen Melvyl**, why do we spend our time in harvesting, loading (some campuses are doing this manually), and maintaining them in our local ILS? Is this a good *return* on *investment* (ROI) model?

3. CAMCIG members agreed that the original cataloging component of the pilot seems to be working fine.

Next Step:

All campuses agreed we should continue with the Shared Record Creation plan divided original cataloging production. The local processes associated with this work flow seemed to work fine. UCSD can share its macro if other campuses are interested.

4. CAMCIG agreed to revisit/reassess CalDocs cataloging to see what we have learned, and what might need to be changed. Might there be other agencies that should be added to the list, or additional campuses that would be willing to assist?

Next Step:

Continue working with the current pilot project arrangement until August 2009, at which point, review what further data/comments we have from selectors, catalogers, Shared Cataloging Program staff and GILS.

Appendix A (Agency Assignments)

Campus Commitments for Original Record Creation

Campus	**California State Agency**
UC Berkeley	Energy Commission
State Water Resources Control Board	
California Policy Research Center	
Coastal Commission	
Franchise Tax Board	
Dept. of Water Resources	
Little Hoover Commission	
UC Davis	Senate Office of Research
Legislative Analyst's Office	
CALFED Bay-Delta Program	
Dept. of Food & Agriculture	

Dept. of Education	
Office of Environmental Health Hazard Assessment	
UC Irvine	Demographic Research Unit
Department of Finance	
Secretary of State	
Division of Communicable Disease Control	
Dept. of Public Health	
Dept. of Industrial Relations, Division of Occupational Safety & Health	
Dept. of Industrial Relations, Division of Labor Statistics and Research	
UCLA	California Labor Market Info
Governor's Office of Planning & Research	
Dept. of Toxic Substances Control	
Dept. of Pesticide Regulation	
Legislative Counsel of California	
Governor's Office of Emergency Services	
Governor's Office on Service & Volunteerism	
UC San Diego	Air Resources Board
California Bureau of State Audits	
California Postsecondary Education Commission	
State Controller's Office	
Board of Equalization	
California Integrated Waste Management Board	